



HAL
open science

Predictive Maintenance in the Industrial Sector: A CRISP-DM Approach for Developing Accurate Machine Failure Prediction Models

Salma Maataoui, Ghita Bencheikh, Ghizlane Bencheikh

► To cite this version:

Salma Maataoui, Ghita Bencheikh, Ghizlane Bencheikh. Predictive Maintenance in the Industrial Sector: A CRISP-DM Approach for Developing Accurate Machine Failure Prediction Models. 2023 Fifth International Conference on Advances in Computational Tools for Engineering Applications (ACTEA), Jul 2023, Zouk Mosbeh, Lebanon. pp.223-227, 10.1109/ACTEA58025.2023.10193983 . hal-04177876

HAL Id: hal-04177876

<https://hal.science/hal-04177876>

Submitted on 11 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Predictive Maintenance in the Industrial Sector: A CRISP-DM Approach for Developing Accurate Machine Failure Prediction Models

1st Salma MAATAOUI
Compute Science and Application
Moulay Ismail University
Meknes, Morocco
salma.maataoui@edu.umi.ac.ma

2nd Ghita BENCHEIKH
LINEACT
Cesi Engineering School
Assat, France
gbencheikh@cesi.fr

3rd Ghizlane BENCHEIKH
Computer Science and Application
Moulay Ismail University
Meknes, Morocco
g.bencheikh@umi.ac.ma

Abstract—In production systems, avoiding repeated failures is crucial for reducing costs and preventing downtime. Industry 4.0 technologies have enabled companies to collect and analyze real-time data from machines, which helps in identifying and preventing potential problems. By using metrics like MTBF and MTTR and analyzing past failures, we can develop predictive models to prevent future failures. This paper explores the use of CRISP-DM methodology in the industrial sector to ensure the accurate prediction of machine failures. Specifically, we examine the application of this methodology in developing predictive models for cutting machines. The results demonstrate that CRISP-DM methodology is effective in developing models that can accurately predict potential failures and prevent them from occurring. The findings have implications for companies looking to implement predictive maintenance strategies in their production systems, highlighting the importance of using data-driven approaches to improve reliability and reduce downtime. Overall, our study highlights the importance of leveraging industry 4.0 technologies and CRISP-DM methodology for optimal performance of production systems in the industrial sector.

Index Terms—Predictive maintenance, Data driven model, CRISP-DM, Industry 4.0.

I. INTRODUCTION

Predictive maintenance (PdM) has become an indispensable tool for managing machinery health in various industries due to the advent of Industry 4.0 and the widespread use of advanced technologies like intelligent automation, machine learning, and artificial intelligence [1], [2]. PdM is even regarded as the driving force behind the fourth industrial revolution [3], and considered as a high priority topic in industry [4]. As modern machines become more complex and the demand for optimal performance increases, PdM has emerged as a proactive maintenance approach that aims to identify and address potential equipment issues before they become major problems [5]. PdM leverages data analytics, sensors, and other cutting-edge technologies to enable companies to optimize their maintenance schedules, minimize downtime, and reduce cost and duration associated with repairs and replacements. Furthermore, PdM helps companies avoid costly shutdowns, enhance safety, and maintain quality standards [6]–[8]. PdM is gaining even more attention from researchers since

artificial intelligence (AI) is involved, offering more evolutionary perspectives [9]. The incorporation of Machine Learning (ML) has transformed decision-making for both individuals and organizations, enabling it to identify illnesses, enhance productivity, optimize transport routes, predict weather, detect fraudulent activities, and much more [10]. By allowing machines to learn from examples and experiences, ML uncovers hidden insights in data [11]. Through data analysis, identifying patterns, and developing predictive logic, machines can predict outcomes without any explicit programming [11], [12]. PdM can benefit greatly from the application of ML techniques in several ways:

- Improved accuracy: ML algorithms can learn from historical data and make predictions with a high level of accuracy. This can help organizations to reduce the risk of unplanned downtime and improve the reliability of their equipment.
- Real-time monitoring: ML algorithms can be applied to real-time sensor data to detect anomalies and patterns that indicate a need for maintenance. This can help organizations to address maintenance issues before they become more serious and lead to downtime.
- Cost savings: By predicting maintenance needs more accurately, organizations can reduce their maintenance costs by avoiding unnecessary maintenance or replacing equipment before it fails.
- Improved asset management: Machine learning can help organizations to better manage their assets by identifying trends and patterns that can help to optimize maintenance schedules, reduce the risk of equipment failure, and extend the lifespan of their equipment.

Over the past few decades, there has been a significant increase in the use of data mining to aid decision-making [13] in manufacturing industries. Kumar and al. propose in their paper referenced [14] a review on the current state of data mining technique, focusing on modern manufacturing methods. Data mining, also known as knowledge discovery in databases, is the process of discovering patterns and knowl-

edge from large datasets. To ensure consistent outcomes from data mining projects, organizations use standardized processes such as Knowledge Discovery in Databases (KDD), Sample, Explore, Modify, Model, and Assess (SEMMA), and Cross-Industry Process for Data Mining (CRISP-DM) for managing data mining projects [15], [16]. However, not every process is suitable for all machine learning objectives [17]. Several studies have been conducted to compare these methods and highlight their strengths and limitations [16]–[18]. Based on these studies, we have selected the CRISP-DM method.

CRISP-DM is widely used in various domains, mostly for finance [15], [19], healthcare [20], [21] and marketing [22], [23]. A bibliographic study of the different methods and application areas of the CRISP-DM method is presented in the article [24]. CRISP-DM has also been successfully applied in manufacturing context [25]. The authors of the papers [26]–[28] have developed a framework to introduce big data analytic into manufacturing systems. Tripathi and al. [29] provide a detailed review of CRISP-DM in industries. However, very few studies use this approach for purpose to predict failures in manufacturing systems [30].

In this study, we propose an application of the generic deep learning CRISP-DM to quantify the current and predict the future degradation of machine by means of health indicators. The approach is tested on a real case study of a cutting machine that operates continuously for 8 hours a day generating a 12 months data retrieved directly from the Cutting Assembly Optimization system (CAO) of the machine.

This paper is organized into four sections starting by an introduction. In section 2, the CRISP-DM method is elaborated. Section 3 presents the case study, business need, data acquisition and pre-processing, and the application of the machine-learning model. Finally, section 4 summarizes the paper’s major findings.

II. BACKGROUND TO CRISP-DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is widely recognized as the industry-standard process model for implementing data mining projects [15], [24]. The authors of the papers [26], [28] aimed to enhance the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is a publicly available standard for carrying out data mining projects. This framework comprises six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, with a particular focus on the first two stages (Fig. 1). The Business Understanding stage aims to comprehend the project’s objectives and requirements from a business perspective, while the Data Understanding stage involves data collection [31]. To bridge the gap between Business Understanding and Data Understanding, the authors introduced intermediate steps. These steps involve transforming business objectives into technical tasks, selecting the data required to accomplish these tasks, and identifying suitable measurement equipment and methodology. These additional steps establish a direct connection between organizational goals and the technical implementation of PdM,

and serve as examples of strategic maintenance implementation in a company.

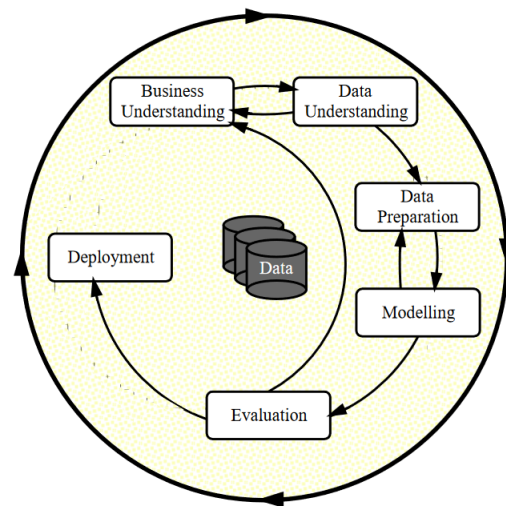


Fig. 1. The CRISP-DM process [15]

A. Description of CRISP-DM steps

The guide of CRISP-DM [31] describes the main idea, tasks and output of these phases shortly, below the summary:

- **Business understanding:** This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
- **Data understanding:** During this phase, data is collected, examined, and explored to identify potential data quality issues, interesting subsets, and possible hypotheses.
- **Data preparation:** Involves building the final data-set for modeling by selecting tables, records, attributes, cleansing and transforming data to suit modeling tools. This data preparation is likely to occur multiple times during the process.
- **Modeling:** Selects and applies various modeling techniques with calibrated parameters to achieve optimal values. Some modeling techniques may require specific data forms, hence the re-execution of the data preparation phase.
- **Evaluation:** Ensures that the constructed models fulfill intended business objectives and comprehensively reviews the previous steps. One of the key objectives is to determine whether any crucial business issues may have been overlooked during the modeling phase.
- **Deployment:** Incorporates the model into decision-making processes, ranging from generating a report to implementing a repeatable data mining process, depending on the project’s requirements.

III. NUMERICAL EXPERIMENT

A. Cutting machine description

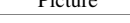



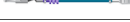

The Crimp Center 64 is a state-of-the-art crimping machine with up to four processing stations.



Fig. 2. Cutting machine Crimp Center 64

The cutting machine is capable of performing high-quality crimping, sealing, twisting and tinning of wires ranging from 0.13 to 6 mm² (26 - 10 AWG) at an impressive maximum feed rate of 12 m/s correspondent to 39.4 ft/s (table I). It boasts exceptional performance, allowing for fast changeovers, high productivity and short set-up times. The machine is user-friendly, equipped with a modern software interface and a touchscreen for easy operation. It also has easy network integration, making it accessible to a range of users. A wide range of accessories and options are available, giving it versatility and adaptability to various applications. The Crimp Center 64 is built for durability, with dynamic and powerful servo drives combined with an intelligent control system that ensures high production rates to meet even the most demanding schedules (Fig. 2). All data, including wire data, crimp data, and seal data, can be saved and retrieved for future use.

TABLE I
PRODUCT TYPES

Picture	Product type
	Partial strip both ends
	2-Step strip both ends
	Crimp to crimp
	Crimp to crimp (closed barrel terminals)
	Crimp to seal
	Seal to seal

At a maximum feed rate of 12 m/s (39.4 ft/s), the Crimp Center 64 allows high-quality crimping, sealing, and tinning of wires from 0.13 to 6 mm² (26 - 10 AWG) at maximum productivity (Table 1).

B. Application

This section covers the implementation of the CRISP-DM methodology to predict machine failures in a wire cutting machine that operates for 8 hours a day. The methodology is implemented on an Intel Core i5 machine with 8GB of RAM and programmed in Python 3.10.

1) *Business understanding*: The project aims to enhance the maintenance KPIs of a manufacturing plant by focusing on two primary objectives.

- to increase the Mean Time Between Failure (MTBF) of a wire cutting machine.
- to achieve a shorter Mean Time To Repair (MTTR).

The wire cutting process involved several operations, including receiving wire spools, cutting wires, performing crimping operations, and conducting quality checks.

- **As-Is State**: Business defined the root of low Overall Equipment Effectiveness (OEE) is the availability of cutting machines.
- **To Be State**: Machine Learning based model that determines the failures.
- **Success**: Reducing the downtime of cutting machines by 50%, the first 6 months and 100%, the second 6 months.
- **Entity**: Improve Mean Time Between failures (MTBF) and Mean Time To Repair (MTTR).
- **Evaluation Metric**: Accuracy.

2) *Data Understanding and Data Preparation*: The maintenance interventions were comprehensively recorded directly in the Cutting and Assembly Optimization system (CAO). This resulted in the creation of an analyzable database that is now available for the necessary analyses. The database includes information such as the machine type, number of goods produced, and downtime (measured by mean time between failures (MTBF) and mean time to repair (MTTR). All historical data is systematically recorded in the system. This study specifically pertains to a cutting machine during the year 2022, from January to December. The dataset contains 549 raw and 20 columns. The MTBF column becomes our target variable. Fig. 4 presents the data visualization.

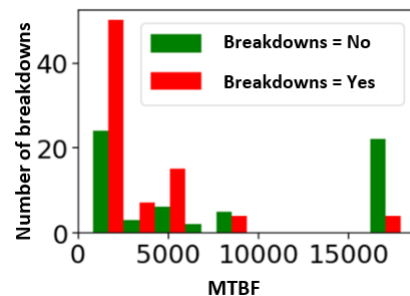


Fig. 3. Visualization of target variable

3) *Modeling*: Selecting the "best model" is a critical step in machine learning, as it determines the accuracy and efficiency of the outcome. This process involves evaluating various

models and selecting the one that performs the best in terms of predictive power, generalization ability, and computational efficiency. A well-informed and careful selection of the best model can greatly enhance the success of the machine-learning task. We have conducted 3 tests in the initial dataset using 3 different methods, respectively Random Forest, Decision Tree, and K-Neighbors, and obtained the results presented in Table II.

TABLE II
MODEL COMPARISON IN THE INITIAL DATASET

Model	Accuracy
Random forest	80%
Decision tree	79%
K-Neighbors	53%

It is evident from the evaluation metrics that the Random Forest model outperforms all other models with perfect scores. On the other hand, the Decision Tree model performs less effectively than the ensemble method. The K-Nearest Neighbors model has the lowest performance, leading to the conclusion that clustering methods might not be the best approach for identifying maintenance failures.

Therefore, we will use the Random Forest Classifier model for classification. This machine-learning algorithm combines multiple decision trees to perform classification. It consists of creating a large number of decision trees and using their collective predictions to classify data accurately [32]. The output of each tree is combined through a voting process to produce the final prediction.

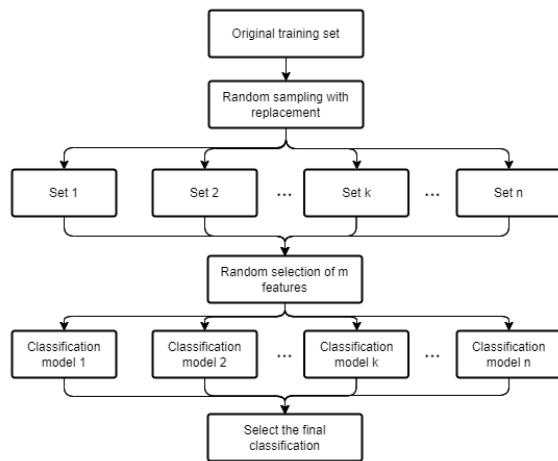


Fig. 4. Flowchart of the Random Forest model

4) *Evaluation and Deployment*: The final work package focused on evaluating the results in a small series of experiments. After preparing the datasets for training, we separated the target from the features and split the data into two parts: 80% for training and 20% for validation. To evaluate the accuracy of our model, we conducted an accuracy test, the results of which are shown in Fig. 5. An epoch in machine learning

refers to a single iteration of the entire training dataset passing through the algorithm during the training process. The number of epochs is a crucial hyper parameter that specifies how many times the entire training dataset will pass through the learning process of the algorithm.

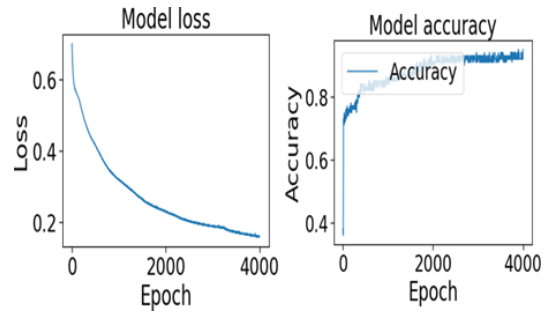


Fig. 5. Accuracy and loss of the model

The test conducted to measure the performance of the model using a dataset of samples showed excellent results, with an accuracy score of 0.8276 indicating a strong performance.

Hyperparameters are parameters that are not learned directly by estimators, but are passed to the estimator class constructor in scikit-learn. Examples include C, kernel, and gamma for Support Vector Classifier, and alpha for Lasso. To achieve the best cross-validation score, it is recommended to search for the optimal hyperparameters. To determine the best parameter setting for our model, we utilize an optimization technique called Grid Search. This method involves testing a range of parameters and comparing their performance (as shown in Figure 6).

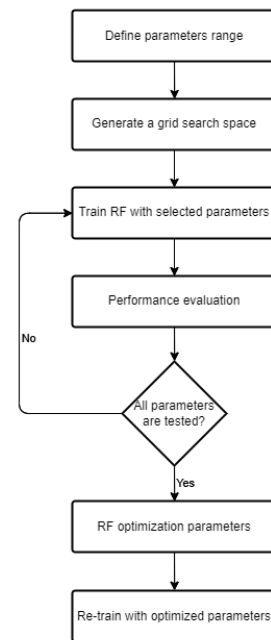


Fig. 6. Flowchart of the Grid Search algorithm

After the retaining of the model, we were able to achieve a higher model score of 0.862.

IV. CONCLUSION

This paper presents the application of the CRISP-DM framework in a study case aimed at improving the key performance indicators (KPIs) of a manufacturing plant. The process involved various stages, including business understanding, data understanding and preparation, modeling, and evaluation and deployment. The study utilized historical data recorded in the Cutting and Assembly Optimization system (CAO). The selected machine learning algorithm was the Random Forest Classifier model, which achieved an accuracy of 0.8276. The model was further optimized using hyper parameter optimization through Grid Search, resulting in a model score of 0.862. The successful application of CRISP-DM in this study case demonstrates the effectiveness of this framework in data-driven decision-making and problem-solving.

REFERENCES

- [1] T. P. Carvalho, F. A. Soares, R. Vita, R. d. P. Francisco, J. P. Basto, and S. G. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Computers & Industrial Engineering*, vol. 137, p. 106024, 2019.
- [2] A. Bousdekis, D. Apostolou, and G. Mentzas, "Predictive maintenance in the 4th industrial revolution: Benefits, business opportunities, and managerial implications," *IEEE Engineering Management Review*, vol. 48, no. 1, pp. 57–62, 2019.
- [3] M. Achouch, M. Dimitrova, K. Ziane, S. Sattarpanah Karganroudi, R. Dhouib, H. Ibrahim, and M. Adda, "On predictive maintenance in industry 4.0: Overview, models, and challenges," *Applied Sciences*, vol. 12, no. 16, p. 8081, 2022.
- [4] S. Zhai, B. Gehring, and G. Reinhart, "Enabling predictive maintenance integrated production scheduling by operation-specific health prognostics with generative deep learning," *Journal of Manufacturing Systems*, vol. 61, pp. 830–855, 2021.
- [5] R. K. Mobley, *An introduction to predictive maintenance*. Elsevier, 2002.
- [6] J. Lindström, H. Larsson, M. Jonsson, and E. Lejon, "Towards intelligent and sustainable production: combining and integrating online predictive maintenance and continuous quality control," *Procedia CIRP*, vol. 63, pp. 443–448, 2017.
- [7] S. Selcuk, "Predictive maintenance, its implementation and latest trends," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 231, no. 9, pp. 1670–1679, 2017.
- [8] J. Dalzochio, R. Kunst, E. Pignaton, A. Binotto, S. Sanyal, J. Favilla, and J. Barbosa, "Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges," *Computers in Industry*, vol. 123, p. 103298, 2020.
- [9] K. T. Nguyen, K. Medjaher, and D. T. Tran, "A review of artificial intelligence methods for engineering prognostics and health management with implementation guidelines," *Artificial Intelligence Review*, pp. 1–51, 2022.
- [10] T. Hong, Z. Wang, X. Luo, and W. Zhang, "State-of-the-art on research and applications of machine learning in the building life cycle," *Energy and Buildings*, vol. 212, p. 109831, 2020.
- [11] S. Thrun and L. Pratt, *Learning to learn*. Springer Science & Business Media, 2012.
- [12] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.
- [13] S. Tufféry, *Data mining and statistics for decision making*. John Wiley & Sons, 2011.
- [14] S. Sahoo, S. Kumar, M. Z. Abedin, W. M. Lim, and S. K. Jakhar, "Deep learning applications in manufacturing operations: a review of trends and ways forward," *Journal of Enterprise Information Management*, vol. 36, no. 1, pp. 221–251, 2023.
- [15] V. Plotnikova, M. Dumas, and F. P. Milani, "Applying the crisp-dm data mining process in the financial services industry: Elicitation of adaptation requirements," *Data & Knowledge Engineering*, vol. 139, p. 102013, 2022.
- [16] U. Shafique and H. Qaiser, "A comparative study of data mining process models (kdd, crisp-dm and semma)," *International Journal of Innovation and Scientific Research*, vol. 12, no. 1, pp. 217–222, 2014.
- [17] A. Däderman and S. Rosander, "Evaluating frameworks for implementing machine learning in signal processing: A comparative study of crisp-dm, semma and kdd," 2018.
- [18] A. Azevedo and M. F. Santos, "Kdd, semma and crisp-dm: a parallel overview," *IADS-DM*, 2008.
- [19] B. C. da Rocha and R. T. de Sousa Junior, "Identifying bank frauds using crisp-dm and decision trees," *International Journal of Computer Science and Information Technology*, vol. 2, no. 5, pp. 162–169, 2010.
- [20] D. Asamoah and R. Sharda, "Adapting crisp-dm process for social network analytics: Application to healthcare," *Twenty-first Americas Conference on Information Systems*, 2015.
- [21] N. Azadeh-Fard, F. M. Megahed, and F. Pakdil, "Variations of length of stay: A case study using control charts in the crisp-dm framework," *International Journal of Six Sigma and Competitive Advantage*, vol. 11, no. 2-3, pp. 204–225, 2019.
- [22] S. Moro, R. Laureano, and P. Cortez, "Using data mining for bank direct marketing: An application of the crisp-dm methodology," 2011.
- [23] S. Peker and Ö. Kart, "Transactional data-based customer segmentation applying crisp-dm methodology: A systematic review," *Journal of Data, Information and Management*, pp. 1–21, 2023.
- [24] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.
- [25] O. J. Fisher, N. J. Watson, J. E. Escrig, R. Witt, L. Porcu, D. Bacon, M. Rigley, and R. L. Gomes, "Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems," *Computers & Chemical Engineering*, vol. 140, p. 106881, 2020.
- [26] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "Dmme: Data mining methodology for engineering applications—a holistic extension to the crisp-dm model," *Procedia Cirp*, vol. 79, pp. 403–408, 2019.
- [27] D. Kozjek, R. Vrabič, B. Rihtaršič, N. Lavrač, and P. Butala, "Advancing manufacturing systems with big-data analytics: A conceptual framework," *International Journal of Computer Integrated Manufacturing*, vol. 33, no. 2, pp. 169–188, 2020.
- [28] H. Wiemer, L. Drowatzky, and S. Ihlenfeldt, "Data mining methodology for engineering applications (dmme)—a holistic extension to the crisp-dm model," *Applied Sciences*, vol. 9, no. 12, p. 2407, 2019.
- [29] S. Tripathi, D. Muhr, M. Brunner, H. Jodlbauer, M. Dehmer, and F. Emmert-Streib, "Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing," *Frontiers in artificial intelligence*, vol. 4, p. 576892, 2021.
- [30] J. A. Harding, M. Shahbaz, and A. Kusiak, "Data mining in manufacturing: a review," 2006.
- [31] C. Pete, C. Julian, K. Randy, K. Thomas, R. Thomas, S. Colin, and R. Wirth, "Crisp-dm 1.0—step-by-step data mining guide," *Cris. Consort*, p. 76, 2000.
- [32] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," in *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*. Springer, 2012, pp. 246–252.